Google | Health Analysis

HealthMan é responsavel por gerenciar o funcionamento de Hospitais de maneira profissional e otimizada. Atualmente eles estão responsável por um novo Hospital na qual eles não possuem informações importantes para a tomada de decisão por parte da equipe de gerenciamento, por essa razão, contrataram você, um Analista de Dados para analisar os dados públicos do novo Hospital. O hospital na qual estão responsável pertence a ID 26.

Tarefa de Negóios * Quais são as tendencias dos pacientes? * Quais são o fluxo dos pacientes e os casos mais frequentes? * Quais departamentos recebem mais pacientes? * O numero de quartos é o suficiente para as necessidades? * Quais outros insights você consegue descobrir pelos dados? * Quais são as suas recomendações para a equipe de gerenciamento

Descrição de dados

traindata.csv: Arquivo contendo as características relacionadas ao paciente, hospital e tempo de permanência por caso
traindata_dictonary.csv: Arquivo contendo as informações das características no arquivo train

Reconhecimentos

Mais detalhes podem ser encontrados no site Analytics Vidhya, que conduziu o hackathon.
https://datahack.analyticsvidhya.com/contest/janatahack-healthcare-analytics-ii/#ProblemStatement

Instalações de Bibliotecas e Exportação de Dados

```
install.packages('tidyverse')
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
install.packages('ggplot2')
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
install.packages('dplyr')
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
```

```
## v ggplot2 3.4.0      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.5.0
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(dplyr)
```

```
# Características relacionadas ao paciente, hospital e tempo de permanência por caso
df <- filter(read_csv('train_data.csv'), Hospital_code == 26)
```

```
## Rows: 318438 Columns: 18
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (9): Hospital_type_code, Hospital_region_code, Department, Ward_Type, Wa...
## dbl (9): case_id, Hospital_code, City_Code_Hospital, Available Extra Rooms i...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# Dicionário
dic <- read_csv('train_data_dictionary.csv')
```

```
## Rows: 18 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): Column, Description
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Limpeza de Dados

- Removendo os espaços nos nomes das colunas
- Removendo a coluna desnecessário do código do hospital
- Convertendo as colunas ID como caractere para que não ocorra operações
- Encurtando a visualização com mais de 100 dias

```
# Removendo os espaços nas colunas e a id do hospitel
colnames(df) <- c('case_id','Hospital_code','Hospital_type_code','City_Code_Hospital','Hospital_region_
                  'Available_Extra_Rooms_in_Hospital','Department','Ward_Type','Ward_Facility_Code',
                  'Bed_Grade','patientid','City_Code_Patient','Type_of_Admission','Severity_of_Illness'
                  'Visitors_with_Patient','Age','Admission_Deposit','Stay')
df <- select(df, -Hospital_code)
```

```
unique(df$Age)
```

```
##  [1] "51-60"  "71-80"  "31-40"  "41-50"  "61-70"  "21-30"  "81-90"  "11-20"
##  [9] "91-100" "0-10"
```

```
#filter(df, Stay == 'More than 100 Days')$Stay <- filter(df, Stay == 'More than 100 Days')$Stay
df$Stay <- replace(df$Stay, df$Stay == 'More than 100 Days', '+100')

summary(select(df, Hospital_type_code, Available_Extra_Rooms_in_Hospital, Bed_Grade, Visitors_with_Pati
```

```
##  Hospital_type_code Available_Extra_Rooms_in_Hospital   Bed_Grade
##  Length:33076       Min.   : 0.000                   Min.   :1.0
##  Class :character   1st Qu.: 2.000                   1st Qu.:2.0
##  Mode  :character   Median : 3.000                   Median :3.0
##                     Mean   : 3.296                   Mean   :2.6
##                     3rd Qu.: 4.000                   3rd Qu.:3.0
##                     Max.   :21.000                   Max.   :4.0
##  Visitors_with_Patient
##  Min.   : 1.000
##  1st Qu.: 2.000
##  Median : 3.000
##  Mean   : 3.333
##  3rd Qu.: 4.000
```

```
##  Max.   :32.000
```

```
# Tratar as IDs como única, para não haver operações entre elas
colnames(df)
```

```
##  [1] "case_id"                          "Hospital_type_code"
##  [3] "City_Code_Hospital"               "Hospital_region_code"
##  [5] "Available_Extra_Rooms_in_Hospital" "Department"
##  [7] "Ward_Type"                        "Ward_Facility_Code"
##  [9] "Bed_Grade"                        "patientid"
## [11] "City_Code_Patient"                "Type_of_Admission"
## [13] "Severity_of_Illness"              "Visitors_with_Patient"
## [15] "Age"                              "Admission_Deposit"
## [17] "Stay"
```

```
df$patientid <- as.character(df$patientid)
df$patientid <- as.character(df$patientid)
dim(df)
```

```
## [1] 33076    17
```

```
head(df)
```

```
## # A tibble: 6 x 17
##   case_id Hospital_typ~1 City_~2 Hospi~3 Avail~4 Depar~5 Ward_~6 Ward_~7 Bed_G~8
##     <dbl> <chr>            <dbl> <chr>     <dbl> <chr>   <chr>   <chr>     <dbl>
## 1       4 b                    2 Y             2 radiot~ R       D             2
## 2       5 b                    2 Y             2 radiot~ S       D             2
## 3      12 b                    2 Y             4 radiot~ R       D             1
## 4      25 b                    2 Y             4 radiot~ Q       D             1
## 5      27 b                    2 Y             4 anesth~ Q       D             3
## 6      28 b                    2 Y             4 gyneco~ R       D             3
## # ... with 8 more variables: patientid <chr>, City_Code_Patient <dbl>,
## #   Type_of_Admission <chr>, Severity_of_Illness <chr>,
## #   Visitors_with_Patient <dbl>, Age <chr>, Admission_Deposit <dbl>,
## #   Stay <chr>, and abbreviated variable names 1: Hospital_type_code,
## #   2: City_Code_Hospital, 3: Hospital_region_code,
## #   4: Available_Extra_Rooms_in_Hospital, 5: Department, 6: Ward_Type,
## #   7: Ward_Facility_Code, 8: Bed_Grade
```

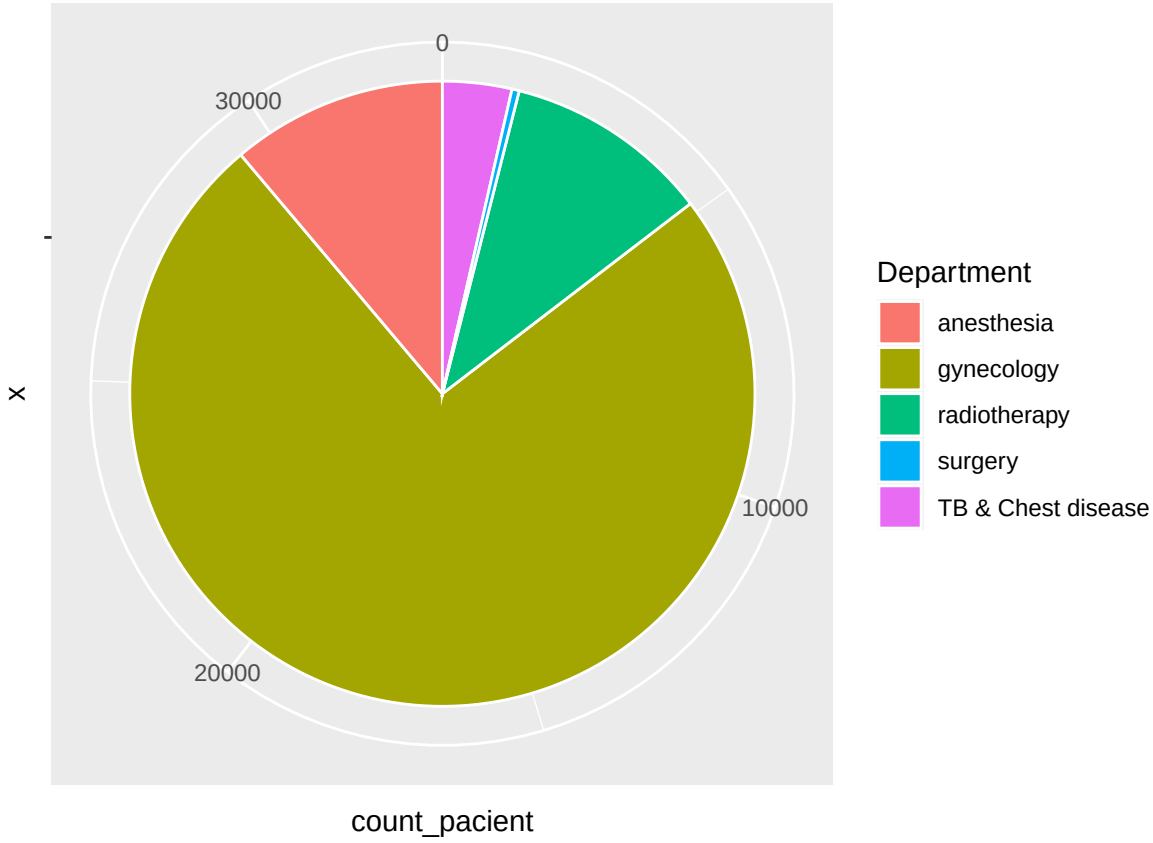
Análise e Visualização

```
fig <- function(width, heigth){options(repr.plot.width = width, repr.plot.height = heigth)}
fig(16,16)

dep <- df %>%
    select(Department, Available_Extra_Rooms_in_Hospital, patientid) %>%
    group_by(Department) %>%
    summarise(mean_rooms = mean(Available_Extra_Rooms_in_Hospital), count_pacient = length(patientid) ) %>%
    arrange(-count_pacient)

dep$frac <- round((dep$count_pacient / sum(dep$count_pacient)) * 100, 1)
dep
```

```
## # A tibble: 5 x 4
##   Department         mean_rooms count_pacient  frac
##   <chr>                   <dbl>         <int> <dbl>
```

```
## 1 gynecology               3.39         24559  74.3
## 2 anesthesia               2.88          3690  11.2
## 3 radiotherapy             3.09          3519  10.6
## 4 TB & Chest disease       3.18          1184   3.6
## 5 surgery                  3.16           124   0.4
```

```
ggplot(data=dep, aes(x='', y=count_pacient, fill=Department)) +
    geom_bar(stat='identity', width=1, color='white') +
    coord_polar('y', start=0)
```




```
ggsave('ggplot01.png')
```

```
## Saving 6.5 x 4.5 in image
```

```
adminission <- df %>%
    select(Type_of_Admission, patientid) %>%
    group_by(Type_of_Admission) %>%
    summarise(count_pacient = length(patientid)) %>%
    arrange(-count_pacient)

adminission$frac <- round((adminission$count_pacient / sum(adminission$count_pacient)) * 100, 1)
adminission
```

```
## # A tibble: 3 x 3
##   Type_of_Admission count_pacient  frac
##   <chr>                    <int> <dbl>
## 1 Trauma                   16695  50.5
## 2 Emergency                10743  32.5
## 3 Urgent                    5638  17
```

```
severity <- df %>%
    select(Severity_of_Illness, patientid) %>%
    group_by(Severity_of_Illness) %>%
    summarise(count_pacient = length(patientid)) %>%
    arrange(-count_pacient)

severity$frac <- round((severity$count_pacient / sum(severity$count_pacient)) * 100, 1)
severity
```

```
## # A tibble: 3 x 3
##   Severity_of_Illness count_pacient  frac
##   <chr>                       <int> <dbl>
## 1 Moderate                    18606  56.3
## 2 Minor                        7780  23.5
## 3 Extreme                      6690  20.2
```

```
adminission_severity <- df %>%
    select(Type_of_Admission, Severity_of_Illness, patientid) %>%
    group_by(Type_of_Admission, Severity_of_Illness) %>%
    dplyr::summarise(count_pacient = length(patientid))
```

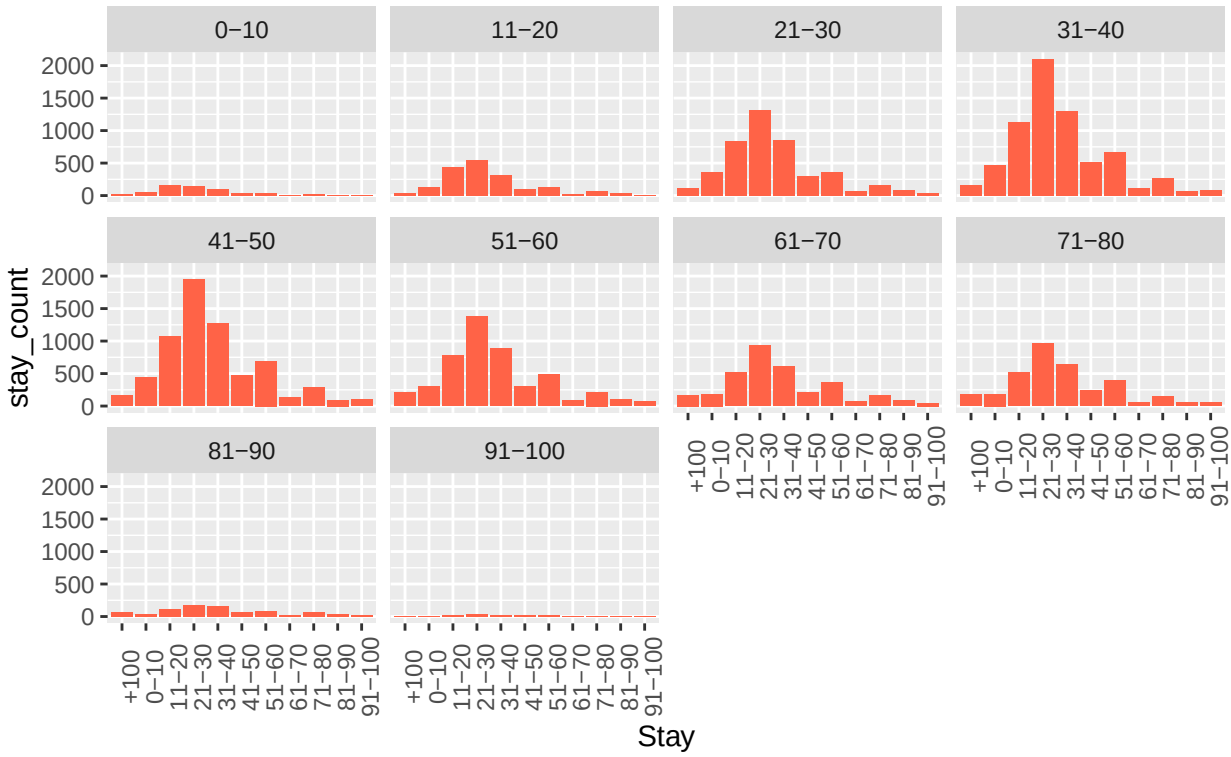
```
## `summarise()` has grouped output by 'Type_of_Admission'. You can override using
## the `.groups` argument.
```

```
adminission_severity$frac <- round((adminission_severity$count_pacient / sum(adminission_severity$count
adminission_severity
```

```
## # A tibble: 9 x 4
## # Groups:   Type_of_Admission [3]
##   Type_of_Admission Severity_of_Illness count_pacient  frac
##   <chr>             <chr>                       <int> <dbl>
## 1 Emergency         Extreme                      1934   5.8
## 2 Emergency         Minor                        3128   9.5
## 3 Emergency         Moderate                     5681  17.2
## 4 Trauma            Extreme                      3688  11.2
## 5 Trauma            Minor                        3256   9.8
## 6 Trauma            Moderate                     9751  29.5
## 7 Urgent            Extreme                      1068   3.2
## 8 Urgent            Minor                        1396   4.2
## 9 Urgent            Moderate                     3174   9.6
```

```
fig(18,9)
df %>%
    select(patientid, Age, Stay) %>%
    group_by(Age, Stay) %>%
    #distinct() %>%
    dplyr::summarise(stay_count = length(Stay)) %>%
    ggplot(aes(x=Stay, y=stay_count)) + geom_bar(stat = 'identity', fill = 'tomato') + facet_wrap('Age'
    theme(axis.text.x = element_text(angle = 90)) +
    labs(title='Frequência por faixa etária e seus dias de hospedagem',
    subtitle='Faixas diária da hospedagem do paciente no hospital')
```

```
## `summarise()` has grouped output by 'Age'. You can override using the `.groups`
## argument.
```




```
ggsave('ggplot02.png')
```

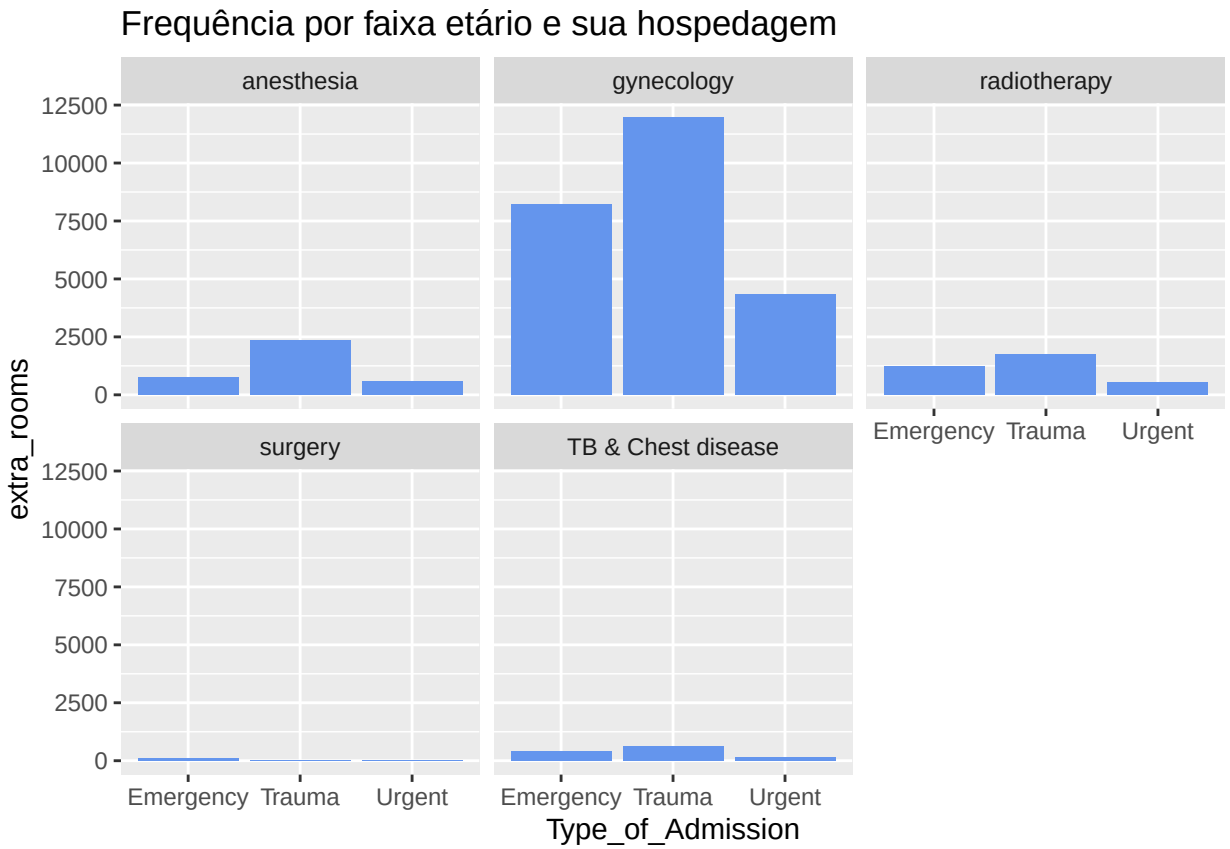
```
## Saving 6.5 x 4.5 in image
```

```
unique(df$Department)
```

```
## [1] "radiotherapy"       "anesthesia"         "gynecology"
## [4] "TB & Chest disease" "surgery"
```

```
fig(16,9)
df %>%  # Condição da cama na enfermaria
    select(Department, Type_of_Admission, Available_Extra_Rooms_in_Hospital, Bed_Grade) %>%
    group_by(Department, Type_of_Admission) %>%
    #distinct() %>%
    dplyr::summarise(extra_rooms = length(Available_Extra_Rooms_in_Hospital), bed_grade = length(Bed_Gra
    ggplot(aes(x=Type_of_Admission, y=extra_rooms)) +
    geom_bar(stat = 'identity', fill = 'cornflowerblue') +
    facet_wrap('Department') +
    labs(title="Frequência por faixa etário e sua hospedagem")
```

```
## `summarise()` has grouped output by 'Department'. You can override using the
## `.groups` argument.
```




```
ggsave('ggplot03.png')
```

```
## Saving 6.5 x 4.5 in image
```

```
summary(df)
```

```
##     case_id       Hospital_type_code City_Code_Hospital Hospital_region_code
##  Min.   :     4   Length:33076       Min.   :2          Length:33076
##  1st Qu.: 74260   Class :character   1st Qu.:2          Class :character
##  Median :153988   Mode  :character   Median :2          Mode  :character
##  Mean   :154725                      Mean   :2
##  3rd Qu.:232482                      3rd Qu.:2
##  Max.   :318433                      Max.   :2
##
##  Available_Extra_Rooms_in_Hospital  Department         Ward_Type
##  Min.   : 0.000                    Length:33076       Length:33076
##  1st Qu.: 2.000                    Class :character   Class :character
##  Median : 3.000                    Mode  :character   Mode  :character
##  Mean   : 3.296
##  3rd Qu.: 4.000
##  Max.   :21.000
##
##  Ward_Facility_Code   Bed_Grade    patientid         City_Code_Patient
##  Length:33076       Min.   :1.0   Length:33076       Min.   : 1.000
##  Class :character   1st Qu.:2.0   Class :character   1st Qu.: 5.000
##  Mode  :character   Median :3.0   Mode  :character   Median : 8.000
##                     Mean   :2.6                      Mean   : 7.705
##                     3rd Qu.:3.0                      3rd Qu.: 8.000
```

```
##                     Max.   :4.0                     Max.   :38.000
##                                                     NA's   :395
##  Type_of_Admission  Severity_of_Illness Visitors_with_Patient
##  Length:33076       Length:33076        Min.   : 1.000
##  Class :character   Class :character    1st Qu.: 2.000
##  Mode  :character   Mode  :character    Median : 3.000
##                                         Mean   : 3.333
##                                         3rd Qu.: 4.000
##                                         Max.   :32.000
##
##      Age            Admission_Deposit     Stay
##  Length:33076       Min.   : 1800     Length:33076
##  Class :character   1st Qu.: 4181     Class :character
##  Mode  :character   Median : 4760     Mode  :character
##                     Mean   : 4898
##                     3rd Qu.: 5467
##                     Max.   :10211
##
```

```
colnames(df)
```

```
##  [1] "case_id"                           "Hospital_type_code"
##  [3] "City_Code_Hospital"                "Hospital_region_code"
##  [5] "Available_Extra_Rooms_in_Hospital" "Department"
##  [7] "Ward_Type"                         "Ward_Facility_Code"
##  [9] "Bed_Grade"                         "patientid"
## [11] "City_Code_Patient"                 "Type_of_Admission"
## [13] "Severity_of_Illness"               "Visitors_with_Patient"
## [15] "Age"                               "Admission_Deposit"
## [17] "Stay"
```

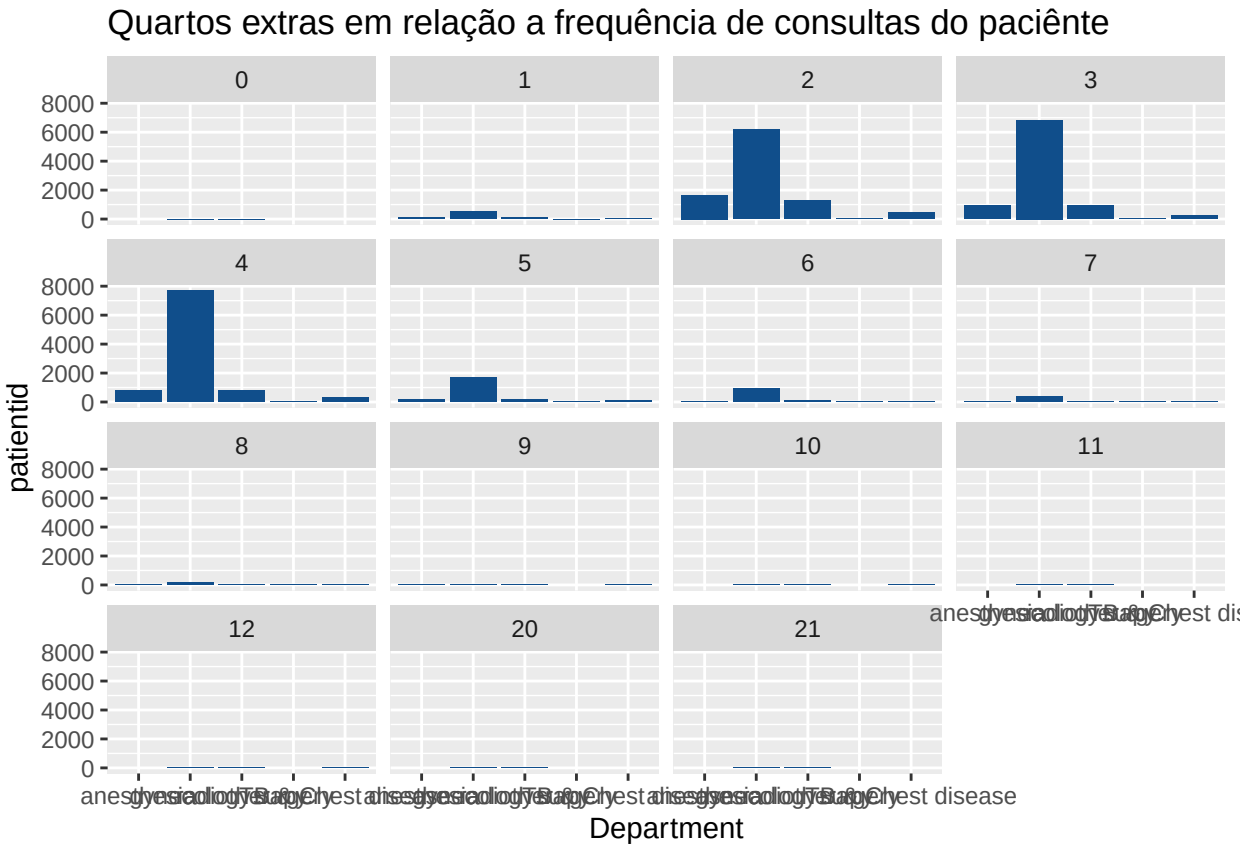
```
head(df)
```

```
## # A tibble: 6 x 17
##   case_id Hospital_typ~1 City_~2 Hospi~3 Avail~4 Depar~5 Ward_~6 Ward_~7 Bed_G~8
##     <dbl> <chr>            <dbl> <chr>     <dbl> <chr>   <chr>   <chr>     <dbl>
## 1       4 b                    2 Y             2 radiot~ R       D             2
## 2       5 b                    2 Y             2 radiot~ S       D             2
## 3      12 b                    2 Y             4 radiot~ R       D             1
## 4      25 b                    2 Y             4 radiot~ Q       D             1
## 5      27 b                    2 Y             4 anesth~ Q       D             3
## 6      28 b                    2 Y             4 gyneco~ R       D             3
## # ... with 8 more variables: patientid <chr>, City_Code_Patient <dbl>,
## #   Type_of_Admission <chr>, Severity_of_Illness <chr>,
## #   Visitors_with_Patient <dbl>, Age <chr>, Admission_Deposit <dbl>,
## #   Stay <chr>, and abbreviated variable names 1: Hospital_type_code,
## #   2: City_Code_Hospital, 3: Hospital_region_code,
## #   4: Available_Extra_Rooms_in_Hospital, 5: Department, 6: Ward_Type,
## #   7: Ward_Facility_Code, 8: Bed_Grade
```

```
fig(15,9)
df %>%
    select(Available_Extra_Rooms_in_Hospital, Department, patientid) %>%
    group_by(Available_Extra_Rooms_in_Hospital, Department) %>%
    summarise(patientid = length(patientid)) %>%
    ggplot(aes(x=Department, y=patientid)) +
```

```
    geom_bar(stat = 'identity', fill = 'dodgerblue4') +
    facet_wrap('Available_Extra_Rooms_in_Hospital') +
    labs(title='Quartos extras em relação a frequência de consultas do paciênte')
```

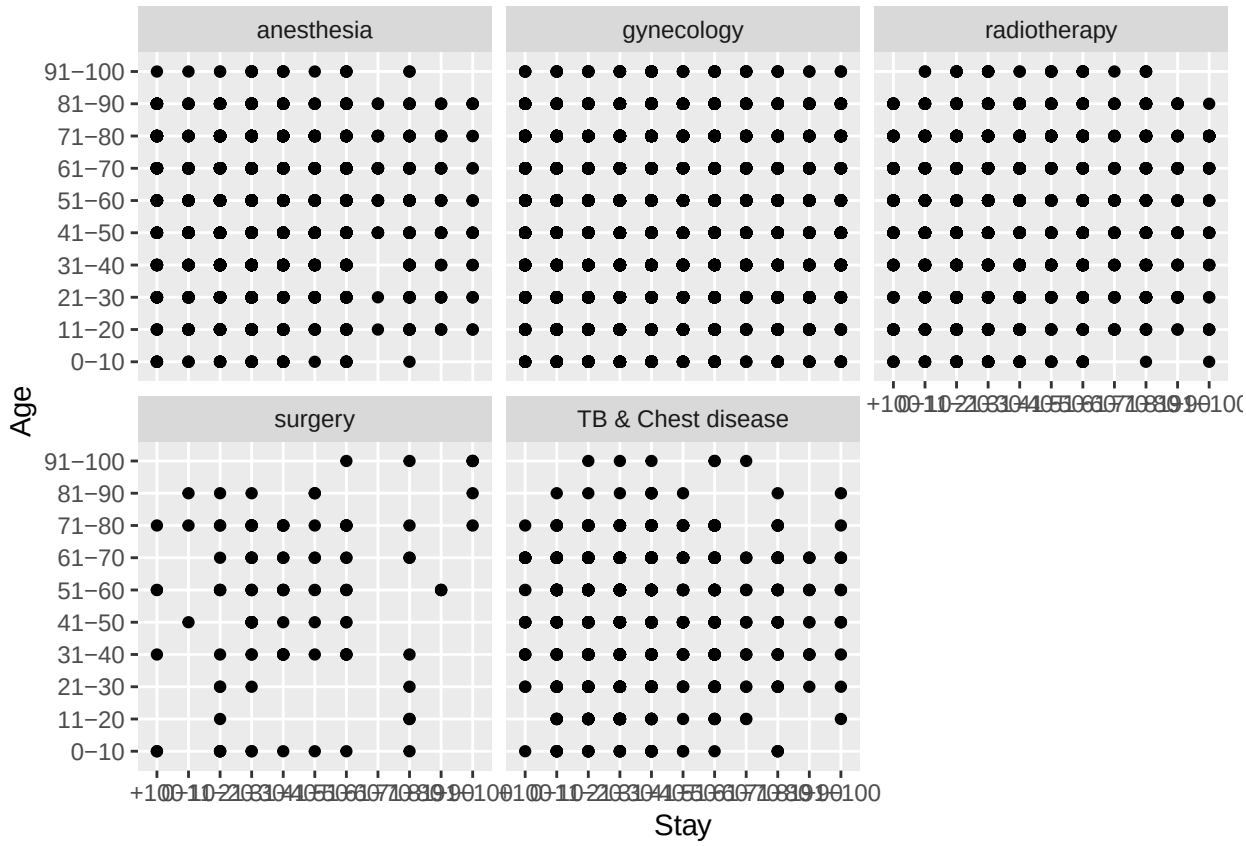
```
## `summarise()` has grouped output by 'Available_Extra_Rooms_in_Hospital'. You
## can override using the `.groups` argument.
```




```
ggsave('ggplot04.png')
```

```
## Saving 6.5 x 4.5 in image
```

```
fig(15,9)
df %>%
    ggplot() + geom_point(mapping=aes(x=Stay, y=Age)) +
    facet_wrap('Department')
```




```
ggsave('ggplot05.png')
```

```
## Saving 6.5 x 4.5 in image
```